Reversing the Norm Effect on Causal Attributions[1]

John Schwenkler and Justin Sytsma

Research in the psychology of causal thinking has frequently revealed effects of normative considerations on causal attributions, where participants tend to assign causality more strongly to agents who violate a norm in bringing about an outcome. Across several experiments, we show that it is possible to reverse this norm effect when the outcome in question is good rather than bad: in these cases, participants assign causality more strongly to a norm-conforming agent than to an agent who violates a norm. We argue that this supports an explanation of the norm effect according to which it is due to a tendency to interpret statements of the form "X caused Y" or "X is the cause of Y" as assigning *responsibility* to X for bringing Y about.

1. Research on the psychology of causal thinking has frequently revealed effects of normative considerations on ordinary causal attributions.[2] If, for example, a professor and an administrative assistant each take one of the last two pens from the secretary's desk, then whichever of them violated a rule against taking pens is identified more strongly as the cause of the secretary's lacking a pen when she needs to record a phone call (Knobe and Fraser 2008). If two wires in a machine both touch a battery and thereby lead the machine to short circuit, the wire that was not supposed to touch the battery is identified more strongly as the cause of the short circuit than the wire that was supposed to touch it (Hitchcock and Knobe 2009). If two people simultaneously log into a computer system leading it to overload and delete an important file, the person who was not supposed to log in is identified more strongly as the cause of the file's being deleted (Reuter et al. 2014). And so on. In each case, causal attributions seem to be

[1] David Rose was significantly involved in this project at an earlier stage, and we thank him for his important contributions. For valuable feedback we also thank two anonymous referees, Josh Knobe, and the members of the Florida State University Philosophical Psychology Group, especially Marshall Thompson and Zina Ward.

[2] By "causal attributions" we specifically mean the use of phrases like "X caused Y" or "X is the cause of Y" (see Sytsma et al. 2019 for further discussion). Some researchers arguably go further than this, asserting that norms matter for causal cognition more generally, evidence is lacking for this more general claim (Danks et al. 2014, Schwenkler and Sievers forthcoming). As such, in this paper we restrict our focus to causal attributions in the minimal sense identified here.

influenced by whether or not the candidate cause violated a norm. Specifically, in these cases the norm-violating entity is judged more strongly as the cause of what happens than the norm-conforming entity.[3] Call this the *norm effect* on causal attributions.

What is the best explanation of this phenomenon? A number of *counterfactual* accounts hold that it is due to the role of normative considerations in counterfactual thinking. For example, on the view of Hitchcock and Knobe (2009) the abnormality of a candidate cause makes it more likely to be taken up in a judgment of the form "If only …" (cf. Kahneman and Tversky 1982), where such a judgment aims to single out what might have been *changed* in order to alter the outcome in question. Thus, for instance, in the first scenario referenced above, if the professor violates a rule in taking a pen but the administrative assistant does not, then the counterfactual expressed by (1) below is more likely to be considered in assessing what happened than is the counterfactual expressed by (2):

(1) If only the professor had not taken a pen, the secretary would have had one when she needed it.

(2) If only the administrative assistant had not taken a pen, the secretary would have had one when she needed it.

While statements (1) and (2) are both true, the fact that the professor's action violated a rule while the administrative assistant did not makes the professor's action more salient, and thus, according to this account, more likely to be considered. This is supposed to explain why the professor is identified more strongly as the cause of the outcome.

This kind of counterfactual account, where normative considerations impact causal attributions via the salience or perceived relevance of specific counterfactuals, is one of the

[3] In each of these cases one entity violates an injunctive norm, while the other does not. Injunctive norms cover both prescriptive norms (what should be done) and proscriptive norms (what should not be done), and while they include distinctively moral norms, they are broader than this, including conventions and etiquette norms, rules and laws, and norms concerning how designed systems are supposed to behave (norms of proper functioning). There is an ongoing debate concerning whether causal attributions are similarly impacted by another type of norm, descriptive or statistical norms, although we will focus on injunctive norms here (but see Hitchcock and Knobe 2009, Sytsma et al. 2012, Livengood et al. 2017, Sytsma forthcoming).

leading accounts in the literature.[4] Not only has this account been used to explain why it would be useful for us to possess a concept of causation that is selective—focusing on only some causal candidates at the expense of others—as opposed to egalitarian (e.g., Hitchcock and Knobe 2009), it also has application to structural equation modeling approaches to causation. One important problem in these approaches is to provide some criteria for determining default and deviant values of variables in a model. As Halpern and Hitchcock (2015) maintain, the kind of normative considerations that feature in counterfactual accounts provide a criterion for setting default and deviant values in models (though see e.g., Blanchard and Schaffer 2017 and Livengood et al. 2017). Moreover, the role of norms in counterfactual judgments has also been brought to bear on the problem of profligate causes that arises in cases of omissions (McGrath 2005). And empirical evidence indicates that counterfactual accounts are indeed relevant to omissions, since causal selection has also been found to follow norm violations in omission cases (Henne et al. 2017; but see Sytsma and Livengood forthcoming).

However, another explanation of these effects is that when participants preferentially respond that one entity caused a certain outcome in studies like those surveyed above, they are expressing something akin to the judgment that the entity is the party *responsible* for the outcome. This basic type of view can be spelled out in several different ways. One division centers on how broad the underlying phenomenon is taken to be. Taking the phenomenon to be relatively narrow, one possibility is that the dominant use of causal attributions (i.e., statements of the form "X caused Y" or "X is the cause of Y") is to express purely descriptive judgments, but that something about the set-up of the empirical studies at issue generates pragmatic considerations that lead participants to instead take the experimenters to be asking for a normative judgment.[5] Taking the phenomenon to be relatively broad, however, an alternative possibility is

[4] See, e.g., Hitchcock and Knobe 2009, Halpern and Hitchcock 2015, Kominsky et al. 2015, Icard et al. 2017, Kominsky and Phillips 2019.

[5] See Samland and Waldmann (2016), Samland et al. (2016).

that the dominant use of causal attributions is not purely descriptive, but has a normative component.[6] On this type of view, statements of the form "X caused Y" or "X is the cause of Y" typically serve to indicate something more than that someone or something contributed to the outcome or brought about the outcome: they also express a normative evaluation, just as statements like "The professor is the one responsible for the problem" or "The professor is the one accountable for the problem" intuitively do. Our purpose in this paper is not to lay out the full range of possibilities here, nor to adjudicate between these competing positions; rather our focus is on the dispute between these views and those according to which the norm effect arises indirectly from the influence of normative considerations on counterfactual thinking.[7] As such we'll refer to these alternative positions jointly as *responsibility accounts*. According to these responsibility accounts, what explains the norm effect is the fact that, at least in these experimental conditions, causal attributions are used by participants to express a normative judgment, i.e., one that assigns responsibility for the outcome in question.

In this paper we critically reconsider one type of evidence that has been put forward in favor of counterfactual accounts and against responsibility accounts. Advocates of counterfactual accounts hold that what matters for causal attributions in cases like those surveyed above is only that a norm was violated, not that the subsequent outcome was bad. Because of this, these accounts yield the prediction that the same effect would be seen if the outcome was instead good. Hitchcock and Knobe make this prediction explicitly: they note that on their view what matters for causal attributions are not judgments about the valence of the effect, but "judgments about whether the candidate cause was itself a norm violation" (2009, p. 603). Since it is only the norm violation that matters, they predict that the valence of the outcome should make no difference to

[6] See, e.g., Sytsma et al. (2012), Livengood et al. (2017), Sytsma et al. (2019), Livengood and Sytsma (2020), Sytsma and Livengood (forthcoming), Sytsma (forthcoming).

[7] These do not exhaust the accounts that have been put forward in the literature. Most notably, Alicke and colleagues have argued that the norm effect is due to people's desire to blame (or praise) the targeted entity biasing their application of what is otherwise a purely descriptive concept (see Alicke 1992, 2000; Alicke et al. 2011; Rose 2017).

the direction of the effect. Specifically, they predict that if a *good* outcome were brought about instead of a *bad* outcome, "the impact of normative considerations should remain unchanged (because people still see that a norm has been violated)" (ibid.).

To test this prediction, Hitchcock and Knobe (2009) gave participants the following *Drug Case*, in which two agents jointly bring about a good outcome:

> An intern is taking care of a patient in a hospital. The intern notices that the patient is having some kidney problems. Recently, the intern read a series of studies about a new drug that can alleviate problems like this one, and he decides to administer the drug in this case.
>
> Before the intern can administer the drug, he needs to get the signature of the pharmacist (to confirm that the hospital has enough in stock) and the signature of the attending doctor (to confirm that the drug is appropriate for this patient). So he sends off requests to both the pharmacist and the attending doctor.
>
> The pharmacist receives the request, checks to see that they have enough in stock, and immediately signs off.
>
> The attending doctor receives the request at the same time and immediately realizes that there are strong reasons to refuse. Although some studies show that the drug can help people with kidney problems, there are also a number of studies showing that the drug can have very dangerous side effects. For this reason, the hospital has a policy forbidding the use of this drug for kidney problems. Despite this policy, the doctor decides to sign off. Since both signatures were received, the patient is administered the drug. As it happens, the patient immediately recovers, and the drug has no adverse effects.

After reading this vignette, participants rated their agreement with a causal attribution concerning either the attending doctor ("The attending doctor's decision caused the patient's recovery") or the pharmacist ("The pharmacist's decision caused the patient's recovery"). Hitchcock and Knobe found evidence of the norm effect: ratings were significantly higher for the attending doctor (M=3.9) than for the pharmacist (M=2.5). And they take this finding to support their counterfactual account over its competitors.

But there are several damning confounds in Hitchcock and Knobe's presentation of this case. First, it could be partly because of the greater degree of responsibility that a doctor, as opposed to a pharmacist, has in respect of a patient's care that the doctor's decision is identified

more strongly as the cause of a patient's recovery. Second, it could also be due to the greater number of words devoted to describing the doctor's decision that the doctor's role is treated as more causally significant. Third, and most importantly for our purposes, while in Hitchcock and Knobe's Drug Case it is clear that the attending doctor violates hospital policy, it is not explicitly stated *why* the doctor does this. The vignette notes that some studies indicate that the drug might help the patient, while others indicate that it might harm the patient. The implication is that the attending doctor is aware of the evidence and is making a difficult decision, presumably guided by the details of the patient's specific case and motivated by a desire to help the patient. In other words, it is not implausible to take the attending doctor to be choosing to uphold a higher norm—the duty of doctors to help their patients—that is in conflict with official policy. Insofar as the attending doctor made a difficult decision in the face of conflicting evidence, while the pharmacist simply checked the stock of the drug, it seems that despite violating hospital policy, the doctor is more deserving of credit for the patient's recovery than is the pharmacist. Following this reasoning, responsibility accounts, like Hitchcock and Knobe's counterfactual account, yield the prediction that the attending doctor will be judged more strongly as the cause of the patient's recovery than the pharmacist. In short, Hitchcock and Knobe's results for the Drug Case do *not* in fact distinguish between the counterfactual and responsibility accounts.

2. To test the hypothesis that the effect observed in the Drug Case reflects a tendency to regard the attending doctor as more deserving of credit for the patient's eventual recovery, we ran a simple study. Participants were given Hitchcock and Knobe's original Drug Case vignette, then asked to rate the following two statements concerning who deserves credit for the patient's recovery using a 7-point scale anchored at 1 with "Strongly disagree," at 4 with "Neither agree nor disagree," and at 7 with "Strongly agree":

The attending doctor deserves credit for the patient's recovery.

The pharmacist deserves credit for the patient's recovery.

The vignette was then repeated on a second page and participants were asked to rate two causal attributions using the same 7-point scale as on the first page:

The attending doctor caused the patient's recovery.

The pharmacist caused the patient's recovery.

Participants were not able to return to the first page. The order of the questions on the first page was randomized and the questions on the second page were presented in that same order.

Participants for each study in this paper were recruited through advertising for a free personality test on Google with the ads displaying in North America.[8] Responses were restricted to participants who indicated that they are native English speakers, 16 years of age or older, with at most minimal training in philosophy, and who had not previously participated in the study.[9] For Study 1, responses were collected from 77 participants who met the restrictions.[10] The results are shown in Figure 1.

We made three predictions concerning Study 1: first, that the effect found by Hitchcock and Knobe would also be found for ratings of how much credit each agent deserves; second, that the effect would replicate for the causal attributions on the second page; and third, that the effect for credit judgments would fully mediate the effect for causal attributions (i.e., the effect for causal attributions would no longer be significant when controlling for the effect on credit judgments). All three predictions were borne out.

---

[8] The personality test was administered after the target questions. One notable benefit of using a "push strategy" like this one (i.e., recruiting participants who were not directly looking to participate in research) is that participants are more likely to be "experimentally naïve" and less likely to be motivated to provide the responses that they think the experimenters are looking for (Haug 2018). Samples collected using the recruitment strategy employed here have been previously compared against samples collected with other methods in replication studies. And the present strategy has been consistently found to generate a diverse sample in terms of geography, socio-economic status, religiosity, political orientation, age, and education. Studies using this strategy have been previously reported in publications including, e.g., Sytsma (2010, 2012), Feltz and Cokely (2011), Murray et al. (2013), Reuter et al. (2014, 2019), Machery et al. (2015), Livengood and Rose (2016), Kim et al. (2016), Sytsma and Ozdemir (2019), Fischer et al. (forthcoming).

[9] Participants were counted as having more than minimal training in philosophy if they had completed an undergraduate major or more advanced studies.

[10] Participants were 67.5% women (three non-binary) and had an average age of 36.3 years (16-90).

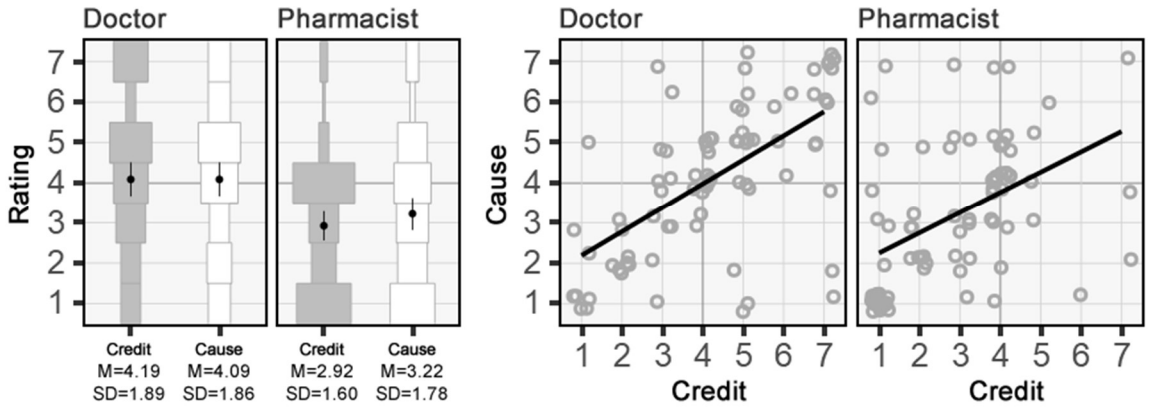

Figure 1: Results for Study 1. Plots on the left show relative percentage of participants selecting each response option, with means and 95% confidence intervals overlaid. Scatterplots on the right show points with jitter and regression lines calculated without jitter.

A two-way ANOVA with *agent* (Doctor, Pharmacist) and *attribution* (Credit, Cause) as within-subjects factors showed a significant main effect for *agent* and no further significant effects, although there was a borderline significant interaction (Table 1). In line with this, planned t-tests revealed the expected effect for both credit judgments, confirming our first prediction, and causal attributions, confirming our second prediction: specifically, in each set of questions ratings were higher for the doctor than for the pharmacist.[11] There was also a strong correlation between the ratings, as is clear from the scatterplots in Figure 1.[12]

Predictor	df_{Num}	df_{Den}	SS_{Num}	SS_{Den}	F	p	η^2_g
(Intercept)	1	76	4007.54	523.21	582.12	.000	.80
agent	1	76	88.39	222.36	30.21	.000	.08
attribution	1	76	0.73	154.02	0.36	.550	.00
agent x *attribution*	1	76	3.12	71.63	3.31	.073	.00

Table 1: Results of ANOVA for Study 1.

To test our third prediction, we performed a Bayesian within-subjects mediation analysis with 10k iterations (Vuorre and Bolger 2018), testing whether participants' credit judgments

[11] Credit: $t(76)=5.54$, $p<.001$, $d=.63$. Cause: $t(76)=3.99$, $p<.001$, $d=.45$.
[12] $r=.57$, $t(152)=8.54$, $p<.001$

mediated the effect on causal attributions. We found that the effect of *agent* on Credit fully mediated the effect on Cause, as seen on the left in Figure 2: the effect is no longer significant when controlling for credit judgments. By contrast, as seen on the right in Figure 2, the reverse analysis shows that while the effect of *agent* on Cause mediated the effect on Credit, it mediated a notably smaller proportion of the effect, and the effect remained significant when controlling for causal attributions.

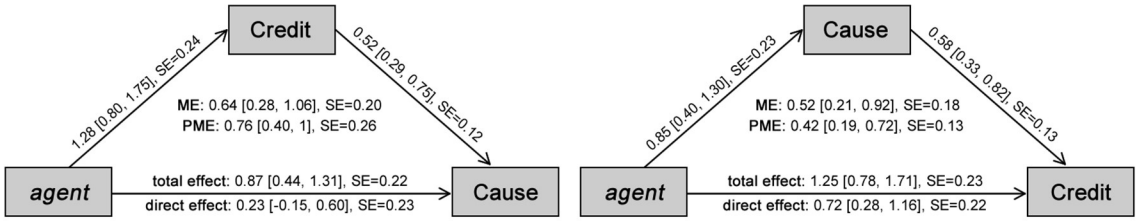

Figure 2: Results for mediation analyses for norm effects in Study 1, showing path diagrams with point estimates (posterior means) of the parameters, with standard errors, and associated 95% credible intervals, as well as the estimated direct effect, mediation effect (ME), and proportion of the effect mediated (PME).

The results of Study 1 suggest that the "norm effect" found by Hitchcock and Knobe (2009) for the Drug Case in fact reflects people's judgments that the attending doctor is more deserving of credit for the patient's recovery than the pharmacist. This is directly in line with the predictions of responsibility accounts. Thus, at worst, the previous findings for the Drug Case do not support counterfactual accounts over responsibility accounts. But arguably the results for Study 1 go further than this, for they raise doubts about whether we are truly dealing with the suggested norm effect in the first place. That is, insofar as the results indicate that the effect for causal attributions is explained by participants' judgments that the doctor is more deserving of credit, it seems that the effect is not specifically due to the doctor's having violated a norm. If anything, following the reasoning above, it would seem that the doctor is more deserving of credit not because she *violated* hospital policy, but because she *followed* a higher norm, using her best judgment in an attempt to help her patient.

3. How might we construct a case that would distinguish more effectively between the predictions of counterfactual accounts and responsibility accounts? Most obviously, the agents in the case need to be such that their roles in bringing about the outcome are more exactly equal, and are described in approximately equal detail. Moreover, it needs to be clear from the description of the case that while the norm-violating agent does *not* deserve credit for bringing about the good outcome, the norm-conforming agent does. In such a case, counterfactual accounts would continue to predict that the norm-violating agent will be identified more strongly as the cause, since it is only the norm violation that matters. By contrast, responsibility accounts predict that in such a case the norm effect will be reversed: the norm-conforming agent will now be identified more strongly as the cause, since it is she who is more deserving of credit.

To test these predictions, we began by adapting the Drug Case to make the agents' motivations clearer. In our modified vignette, two physicians sign off on the request to treat a patient with a certain drug, with one of them clearly violating a norm in doing so while the other clearly does not. In line with the above reasoning, we specified that the norm-violating doctor (Dr. Smith) signs off on the request for a bad reason, with no concern for how this would impact the patient. This is contrasted with a norm-conforming doctor (Dr. Patel) who signs off on the request for a good reason, believing that the drug will help the patient. The revised vignette reads as follows:

> An intern is taking care of a patient in a hospital. The intern notices that the patient is having some kidney problems. The intern knows that a certain drug is often administered for kidney problems like these, and he thinks it might be good to administer the drug in this case.
>
> Before the intern can administer the drug, he needs to get the signatures of the two attending physicians, Dr. Smith and Dr. Patel. So he sends off requests to both of them. Although the drug is often administered in the hospital, hospital policy nonetheless requires attending physicians to carefully consider a patient's file, including information about their condition and medical history, before signing off on the administration of any drug.
>
> Dr. Patel receives the request. He is aware that many studies show that the drug can help people with kidney problems, but that there are also some studies showing that the drug

> can have dangerous side effects depending on the patient's medical history. Dr. Patel reviews the patient's file and realizes that the situation is complicated: the patient has some indicators of increased risk for side effects, but not others. After carefully reviewing the evidence, however, Dr. Patel concludes that the balance of considerations point in favor of administering the drug, and so he decides to sign off on the request.
>
> Dr. Smith receives the request at the same time. Like Dr. Patel, he is aware that many studies show that the drug can help people with kidney problems, but that there are also some studies showing that the drug can have dangerous side effects depending on the patient's medical history. But Dr. Smith does not bother to review the patient's file and, thus, is ignorant of the patient's medical history. Dr. Smith has a close relationship with the pharmaceutical company that manufactures this drug and gets a kick-back every time he approves the drug, so he decides to sign off on the request for that reason.
>
> Since both signatures are received, the patient is administered the drug.
>
> As it happens, the patient's condition immediately improves, and the drug has no adverse side effects.

Our second study followed the first, using a two-page design with participants being asked to rate a pair of statements on each page using the same scale as before. This time, on the first page participants rated the following pair of causal attributions:

> Dr. Patel caused the patient's condition to improve.
>
> Dr. Smith caused the patient's condition to improve.

In addition, after the causal attributions participants were given a check question asking them, "How many attending physicians needed to sign off on the request to administer the drug?" On the second page, participants then rated the following pair of responsibility attributions:

> Dr. Patel is responsible for the patient's condition improving.
>
> Dr. Smith is responsible for the patient's condition improving.

As before, the order of the questions was randomized, the vignette was repeated on the second page, and participants were not able to go back to the first page after proceeding. Responses were

collected from 73 participants who met the restrictions and passed the check question.[13] The results are shown in Figure 3.

As we have seen, counterfactual accounts predict that causal ratings for Dr. Smith will be significantly higher than for Dr. Patel. After all, it is Dr. Patel who conforms with the relevant norms, while Dr. Smith not only violates the hospital's policy, but violates clear ethical norms in signing off on the request for purposes of receiving a kick-back. By contrast, responsibility accounts make the opposite prediction: they predict that causal ratings for Dr. Patel will be significantly higher than for Dr. Smith, since Dr. Patel is more deserving of credit for the patient's condition improving. Further, responsibility accounts predict that we will see the same general effect for responsibility attributions that we expect to find for causal attributions. The results of Study 2 ran counter to the prediction of counterfactual accounts, while both predictions of responsibility accounts were borne out.

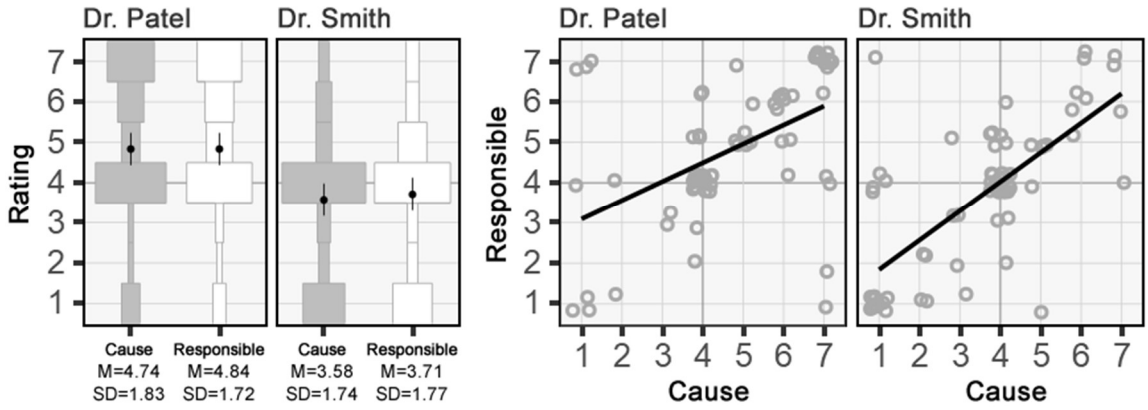

Figure 3: Results for Study 2. Plots on the left show relative percentage of participants selecting each response option, with means and 95% confidence intervals overlaid. Scatterplots on the right show points with jitter and regression lines calculated without jitter.

A two-way ANOVA with *agent* (Patel, Smith) and *attribution* (Cause, Responsible) as within-subjects factors showed a significant main effect for *agent* and no further effects (Table 2).

[13] 2/75 (2.6%) of participants missed the check question. The remaining participants were 73.6% women (one non-binary) and had an average age of 50.2 years (16-84).

The lack of significant effects for *attribution* supports the second prediction of responsibility accounts.[14] Planned t-tests revealed that against the prediction of counterfactual accounts, and in line with that of responsibility accounts, the norm effect was *reversed* for the causal attributions, with ratings being higher for the norm-conforming agent (Dr. Patel) than for the norm-violating agent (Dr. Smith).[15] And, again in line with the prediction of responsibility accounts, the norm effect was also *reversed* for responsibility attributions.[16] Further, there was a strong correlation between ratings for Cause and Responsible, as is clear from the scatterplots in Figure 3.[17]

Predictor	df_{Num}	df_{Den}	SS_{Num}	SS_{Den}	F	p	η^2_g
(Intercept)	1	72	5189.59	471.16	793.05	.000	.85
agent	1	72	95.51	248.24	27.70	.000	.10
attribution	1	72	0.99	97.76	0.73	.396	.00
agent x *attribution*	1	72	0.03	79.72	0.03	.868	.00

Table 2: Results of ANOVA for Study 2.

The findings for Study 2 are readily explained by responsibility accounts but are quite problematic for counterfactual accounts. Most importantly, we find the *reverse* effect of that predicted by counterfactual accounts for causal attributions: participants identified the *norm-conforming* agent more strongly as the cause of the patient's recovery. Following the above discussion, the most natural explanation is that participants were inclined to give Dr. Patel credit

[14] The finding that ratings for causal attributions and responsibility attributions are not statistically significantly distinguishable is in line with previous studies (e.g., Sytsma forthcoming) and, arguably, supports the responsibility view put forward by Sytsma and Livengood over the pragmatic view put forward by Samland and Waldmann and the bias view put forward by Alicke and colleagues. While the former treats the dominant use of causal attributions as expressing a normative concept akin to responsibility, and so expects a close correspondence between causal attributions and responsibility attributions in cases like this, the latter views hold that pragmatic features or bias are leading some participants astray. As such, while these views would predict similar effects for causal attributions and responsibility attributions, they would *not* predict the close correspondence observed here as this would suggest an unrealistically strong pragmatic or bias effect. See Livengood and Sytsma (2020) and Sytsma (forthcoming) for discussion.

[15] $t(72)=4.87$, $p<.001$, $d=.57$

[16] $t(72)=4.32$, $p<.001$, $d=.51$

[17] $r=.64$, $t(144)=10.06$, $p<.001$

for the patient's improvement because he deliberated over the patient's case in the way that he was supposed to do. As such, the reverse effect is in keeping with responsibility accounts.

4. The counterfactual account proposed by Hitchcock and Knobe (2009) has subsequently been developed in a number of papers, including by Kominsky et al. (2015). There they explore a further effect that they term "causal superseding." The details need not concern us here; what is important for present purposes is that Kominsky et al. make a similar prediction to Hitchcock and Knobe concerning the impact of outcome valence. As they write, "from the standpoint of the counterfactual account, the relevant component is the norm violation of the superseding actor, not the valence of the outcome" (2009, p. 199).

In Kominsky et al.'s second experiment they tested a different case—the *Computer Case*—involving a good outcome. The vignette for the relevant condition reads as follows:

> Billy and Suzy work for the same company. They work in different rooms and both of them sometimes need to access the central computer of the company. Nobody at the company is aware that if two people are logged into the central computer at the same time, some spam e-mails containing dangerous viruses are immediately deleted from the central computer.
>
> In order to make sure that one person is always available to answer incoming phone calls, the company issued the following official policy: Suzy is the only one permitted to log into the central computer in the mornings, whereas Billy is the only one permitted to log into the central computer in the afternoons. Billy is not permitted to log into the central computer in the morning.
>
> Today at 9am, Billy and Suzy both log into the central computer at the same time. Immediately, some work e-mails containing dangerous viruses are deleted from the central computer.

On a first page, participants rated a pair of causal attributions on a seven-point scale—one stating that Billy caused the outcome, the other that Suzy caused the outcome. While Kominsky et al. do not report the individual means for this condition, it is clear from their Figure 3 that the norm effect occurred, with ratings being notably higher for Billy than for Suzy.

However, as with Hitchcock and Knobe's original Drug Case, it would be a mistake to use the Computer Case as a way of testing counterfactual accounts against responsibility accounts, since it simply is not clear what responsibility accounts should predict about judgments concerning this case. For since neither Billy nor Suzy knew about the issue with the central computer, neither could have logged in with the *intention* of thereby bringing about the deletion of the emails containing the dangerous viruses. As such, the reasoning applied for the revised Drug Case does not apply to the Computer Case, since Suzy seems not to deserve any *credit* for the unforeseen good outcome of her act. Following the results from the previous study, however, we predicted that the norm effect would again be reversed if it was specified that Suzy acted specifically for the purpose of deleting the dangerous viruses, while Billy acted for an alternative, malicious reason.

To test this prediction, in our third study we gave participants a revised version of the Computer Case in which the motives of both agents are now explicitly noted. The revised vignette reads as follows:

> Billy and Suzy work for the same company. They work in different rooms and both of them sometimes need to access the company's central computer.
>
> As a safety procedure to ensure that important files aren't lost, company policy requires that in order for a file to be deleted from the central computer, two employees must mark the file for deletion. Only some employees are allowed to mark files for deletion. Suzy is allowed to mark files for deletion, but Billy is not allowed to mark files for deletion.
>
> This morning, Billy and Suzy both marked the same file for deletion at the exact same time.
>
> Suzy had looked carefully at the metadata for the file and came to the considered conclusion that the file contained a dangerous computer virus. She marked the file for deletion in order to remove the virus.
>
> Billy had looked at the name of the file, and based on the name concluded that the file contained financial records that would reveal that he's been embezzling money from the company. So Billy marked the file for deletion in order to remove the incriminating evidence.
>
> Since two employees had both marked the file for deletion, the file was immediately deleted from the central computer. As it turns out, the deleted file contained an extremely dangerous computer virus.

The approach of Study 3 followed the first two, using a two-page design with participants being asked to rate a pair of statements on each page using the same scale as before. As in the previous study, on the first page participants rated a pair of causal attributions:

Billy caused the file containing the dangerous virus to be deleted.

Suzy caused the file containing the dangerous virus to be deleted.

After the causal attributions participants were again given a check question: "How many people needed to mark the file for deletion for it to be deleted?" On the second page participants then rated a pair of responsibility attributions:

Billy is responsible for the file containing the dangerous virus being deleted.

Suzy is responsible for file containing the dangerous virus being deleted.

As before, the order of the questions was randomized, the vignette was repeated on the second page, and participants were not able to go back to the first page after proceeding. Responses were collected from 104 participants who met the restrictions and passed the check question.[18] The results are shown in Figure 4.

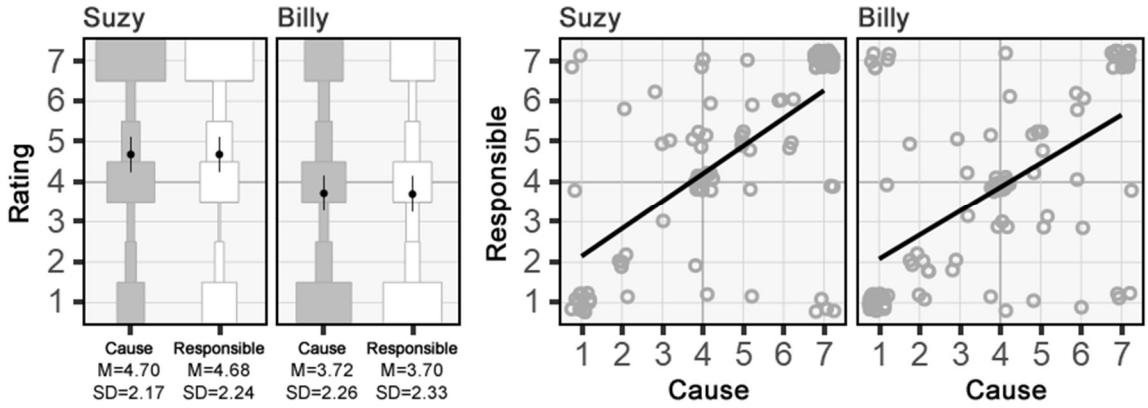

Figure 4: Results for Study 3. Plots on the left show relative percentage of participants selecting each response option, with means and 95% confidence intervals overlaid. Scatterplots on the right show points with jitter and regression lines calculated without jitter.

[18] 17/121 (14.0%) of participants missed the check question. The remaining participants were 75.0% women (one non-binary) and had an average age of 53.3 years (16-86).

As with the previous study, counterfactual accounts predict that causal ratings for Billy should be significantly higher than for Suzy, since Billy violated company policy (in addition to ethical and legal norms in attempting to destroy incriminating evidence), while Suzy did not. By contrast, responsibility accounts again make the opposite prediction: causal ratings for Suzy should be significantly higher than for Billy, since Suzy is more deserving of credit for the dangerous virus being deleted. Further, responsibility accounts predict that we will see the same effect for responsibility attributions that we expect to find for causal attributions. The results of our experiment again ran counter to the prediction of counterfactual accounts, and were in line with those of responsibility accounts instead.

A two-way ANOVA with *agent* (Suzy, Billy) and *attribution* (Cause, Responsible) as within-subjects factors again showed a significant main effect for *agent* and no further effects (Table 3). The lack of significant effects for *attribution* supports the second prediction of responsibility accounts. Planned t-tests revealed that, as predicted by responsibility accounts but not by counterfactual accounts, the norm effect was again *reversed* for the causal attributions, with ratings being higher for the norm-conforming agent (Suzy) than for the norm-violating agent (Billy).[19] And, in line with the second prediction of responsibility accounts, the norm effect was also *reversed* for the responsibility attributions.[20] Further, there was once again a strong correlation between ratings for Cause and Responsible, as is clear from the scatterplots.[21]

Predictor	df_{Num}	df_{Den}	SS_{Num}	SS_{Den}	F	p	η^2_g
(Intercept)	1	103	7344.96	1025.04	738.05	.000	.78
agent	1	103	100.04	660.96	15.59	.000	.05
attribution	1	103	0.04	140.96	0.03	.867	.00
agent x *attribution*	1	103	0.00	258.00	0.00	1.00	.00

Table 3: Results of ANOVA for Study 3.

[19] $t(103)=3.39$, $p<.001$, $d=.33$
[20] $t(103)=3.31$, $p<.001$, $d=.32$
[21] $r=.64$, $t(206)=11.80$, $p<.001$

Study 3 replicated the findings from Study 2 with a second case from the literature. As in the previous study, we found the *reverse* effect of that predicted by counterfactual accounts for causal attributions. This puts further pressure on counterfactual accounts, while again lining up with the predictions of responsibility accounts.

5. Daniel Kahneman (2011, 133) writes that "the proof that you truly understand a pattern of behavior is that you know how to reverse it."[22] In this paper we reversed an important effect found in the literature on the impact of norms on causal attributions. Advocates of counterfactual accounts have taken the occurrence of the norm effect for cases with good outcomes to offer strong support for their accounts while providing evidence against responsibility accounts. Based on insights from responsibility accounts, however, we suggested an alternative explanation of the effect. This explanation was supported by the results of our first study. We then called on this explanation to predict when the effect would be reversed. And, indeed, the reverse effect was found for two different vignettes in our second and third studies. These studies provide strong evidence against counterfactual accounts. And, if Kahneman is right and the proof that you understand an effect is seen in the ability to reverse it, these studies provide equally strong evidence in favor of responsibility accounts.

References

Alicke, M. (1992). "Culpable causation." *Journal of Personality and Social Psychology*, 63: 368–378.

Alicke, M. (2000). "Culpable Control and the Psychology of Blame." *Psychological Bulletin*, 126(4): 556–574.

Alicke, M., Rose, D., and Bloom, D. (2011). "Causation, Norm Violation and Culpable Control." *Journal of Philosophy*, 108: 670-696.

Blanchard, T. and J. Schaffer (2017). "Cause without default." In H. Beebee, P. Menzies, and C. Hitchcock (eds.), "Making a difference," pp. 175-214, Oxford University Press.

[22] See Robinson, Stey, and Alfano (2015) for another application of this dictum.

Danks, D., D. Rose, and E. Machery (2014). "Demoralizing Causation." *Philosophical Studies*, 171: 251–277.

Feltz, A. and E. Cokely (2011). "Individual differences in theory-of-mind judgments: Order effects and side effects." *Philosophical Psychology*, 24(3): 343-355.

Fischer, E., P. Engelhardt, and J. Sytsma (forthcoming). "Inappropriate stereotypical inferences? An adversarial collaboration in experimental ordinary language philosophy." *Synthese*.

Halpern, J. and C. Hitchcock (2015). "Graded Causation and Defaults." *British Journal for the Philosophy of Science*, 66: 413–457.

Haug, M. (2018). "Fast, Cheap, and Unethical? The Interplay of Morality and Methodology in Crowdsourced Survey Research." *Review of Philosophy and Psychology*, 9(2): 363-379.

Henne, P., Á. Pinillos, and F. De Brigard (2017). "Cause by Omission and Norm: Not Watering Plants." *Australasian Journal of Philosophy*, 95(2): 270–283.

Hitchcock, C. and J. Knobe (2009). "Cause and Norm." *The Journal of Philosophy*, 106: 587–612.

Icard, T., J. Kominsky, and J. Knobe (2017). "Normality and Actual Causal Strength." *Cognition*, 161: 80–93.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kahneman, D. and A. Tversky (1982). "The psychology of preferences." *Scientific American*, 246(1): 160–173.

Kim, H., N. Poth, K. Reuter, and J. Sytsma (2016). "Where is your pain? A Cross-cultural Comparison of the Concept of Pain in Americans and South Koreans." *Studia Philosophica Estonica*, 9(1): 136-169.

Knobe, J. and B. Fraser (2008). "Causal judgments and moral judgment: Two experiments." In W. Sinnott-Armstrong (ed.), *Moral Psychology, Volume 2: The Cognitive Science of Morality*, pp. 441–447, Cambridge: MIT Press.

Kominsky, J., J. Phillips, T. Gerstenberg, D. Lagnado, and J. Knobe (2015). "Causal superseding." *Cognition*, 137: 196–209.

Kominsky, J. and J. Phillips (2019). "Immoral Professors and Malfunctioning Tools: Counterfactual Relevance Accounts Explain the Effect of Norm Violations on Causal Selection." *Cognitive Science*, 43(11): e12792.

Livengood, J. and D. Rose (2016). "Experimental Philosophy and Causal Attribution." In J. Sytsma and W. Buckwalter (eds.), *A Companion to Experimental Philosophy*, Wiley Blackwell, 434-449.

Livengood, J. and J. Sytsma (2020). "Actual causation and compositionality." *Philosophy of Science*, 87(1): 43-69.

Livengood, J., J. Sytsma, and D. Rose (2017). "Following the FAD: Folk attributions and theories of actual causation." *Review of Philosophy and Psychology*, 8(2): 274–294.

Machery, E., J. Sytsma, and M. Deutsch (2015). “Speaker’s Reference and Cross-cultural Semantics.” In A. Bianchi (ed.), *On Reference*, Oxford University Press, 62-76.

Murray, D., J. Sytsma, and J. Livengood (2013). “God Knows (But does God Believe?)” *Philosophical Studies*,166: 83-107.

Reuter, K., Kirfel, L., van Riel, R., and Barlassina, L. (2014). “The good, the bad, and the timely: how temporal order and moral judgment influence causal selection.” *Frontiers in Psychology*, 5: 1336.

Reuter, K., D. Phillips, and J. Sytsma (2014). “Hallucinating Pain.” In J. Sytsma (Ed.), *Advances in Experimental Philosophy of Mind*, Bloomsbury, 75-100.

Reuter, K., M. Sienhold, and J. Sytsma (2019). “Putting Pain in its Proper Place.” *Analysis*, 79(1): 72-82.

Robinson, B., P. Stey, and M. Alfano (2015). “Reversing the side-effect effect: the power of salient norms.” *Philosophical Studies*, 172: 177–206.

Rose, D. (2017). “Folk Intuitions of Actual Causation: A Two-pronged Debunking Explanation.” *Philosophical Studies*, 174(5): 1323–1361.

Samland, J. and M. R. Waldmann (2016). “How prescriptive norms influence causal inferences.” *Cognition*, 156: 164–176.

Samland, J., M. Josephs, M. Waldmann, and H. Rakoczy (2016). “The Role of Prescriptive Norms and Knowledge in Children’s and Adults’ Causal Selection.” *Journal of Experimental Psychology: General*, 145(2): 125–130.

Schwenkler, J., and E. T. Sievers (forthcoming). “Cause, ‘Cause’, and Norm.” To appear in P. Willemsen and A. Wiegmann (eds.), *Advances in Experimental Philosophy of Causation*, London: Bloomsbury Press.

Sytsma, J. (2010). “Dennett’s Theory of the Folk Theory of Consciousness.” *Journal of Consciousness Studies*, 17(3-4): 107-130.

Sytsma, J. (2012). “Revisiting the Valence Account.” *Philosophical Topics*, 40(2): 179-198.

Sytsma, J. (forthcoming). “Causation, Responsibility, and Typicality.” *Review of Philosophy and Psychology*.

Sytsma, J., R. Bluhm, P. Willemsen, and K. Reuter (2019). “Causal Attributions and Corpus Analysis.” In E. Fischer and M. Curtis (eds.), *Methodological Advances in Experimental Philosophy*, London: Bloomsbury Press.

Sytsma, J. and J. Livengood (forthcoming). “Causal Attributions and the Trolley Problem.”

Sytsma, J., J. Livengood, and D. Rose (2012). “Two types of typicality: Rethinking the role of statistical typicality in ordinary causal attributions.” *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43: 814–820.

Sytsma, J. and E. Ozdemir (2019). “No Problem: Evidence that the Concept of Phenomenal Consciousness is Not Widespread.” *Journal of Consciousness Studies*, 26(9-10): 241-256.

Vuorre, M. and N. Bolger (2018). “Within-subject mediation analysis for experimental data in cognitive psychology and neuroscience.” *Behavior Research Methods*, 50: 2125–2143.